〈共同研究プロジェクト紹介〉萌芽・発掘型 : 統計と機械学習による日本語史研究 歴史的日本語資料のアノテーションと自動濁点付与

著者	小木曽 智信
雑誌名	国語研プロジェクトレビュー
巻	4
号	2
ページ	144-150
発行年	2013-10
URL	http://doi.org/10.15084/00000743

国語研プロジェクトレビュー
NINJAL Project Review　Vol.4 No.2 pp.144–150（October 2013）

歴史的日本語資料のアノテーションと自動濁点付与

Analysis of Historical Japanese Texts and Automatic *dakuten* Annotation

小木曽 智信（OGISO Toshinobu）

1. はじめに

国立国語研究所では，基幹型プロジェクト「通時コーパスの設計」（近藤泰弘プロジェクトリーダー）を中心として，日本語の歴史を研究することのできるコーパスの構築に取り組んでいる。また，独創・発展型プロジェクト「近代語コーパス設計のための文献言語研究」（田中牧郎プロジェクトリーダー）では，明治時代の『明六雑誌コーパス』[1]（近藤・田中 2012）の構築を行ってきた。こうした歴史的日本語資料のコーパス構築では，単にテキストを機械可読にするだけでなく，資料の構造や原文の状態に関する情報をテキストに付与し，さらに単語情報などを付与することが求められる。こうした情報付与（アノテーション）を行うことで，コーパスを活用した高度な検索や集計，統計的処理が可能になる。これにより，日本語史の研究においてコーパス言語学で培われた手法を応用し，新たな知見をもたらすことが期待される。

多くのテキストに対して正確で均質なタグ付けを行うためには，形態素解析をはじめとするさまざまな機械処理が必要となる。萌芽・発掘型プロジェクト「統計と機械学習による日本語史研究」では，自然言語処理において応用が進んでいる統計的機械学習にもとづいて，歴史的な日本語資料を対象としたアノテーションのための技術開発を行ってきた。また，アノテーションが施されたコーパスを用いて，従来行うことのできなかった統計的手法にもとづく日本語史の研究に取り組んでいる。

本稿では，歴史的日本語資料のアノテーションの流れを俯瞰した上で，アノテーション作業の自動化の試みの一つとして当プロジェクトで開発された濁点の自動付与に関する研究成果を紹介する。最後に，今後の歴史コーパスに期待される高度なアノテーションについて展望する。

2. 歴史的資料のアノテーション

歴史的な資料のコーパスを構築するとき，テキストに対するさまざまな次元でのアノテーションが必要となる。たとえば，一般的な文字セットで表現できない外字，漢字の左右に付される振り仮名・割書などの独特の書式，表題と本文・台詞とト書き・注釈・奥付などの原

[1] http://www.ninjal.ac.jp/corpus_center/cmj/meiroku/

資料の構造などがある。『太陽コーパス』（国立国語研究所編 2005）ではこのレベルまでのアノテーションが行われている。通時コーパスでは，これらの情報は国際的なガイドラインである TEI[2] を参考にしながら XML を用いて人手によってタグ付けされている。

こうして作られた構造化文書に単語情報を付与するには，本文を形態素解析が可能な状態にまで整備する必要がある。たとえば，文の境界のアノテーションもその一つである。歴史的資料では，今日のように句点が必ず用いられるわけではないため，文の区切りの認定は容易ではない。このほか，踊り字で繰り返しが表現された部分や，漢文のように日本語の語順とは異なる順で文字が書かれている部分では前もって読み下しておかなければならない場合がある。こうした処理も原文の状態を保存するために XML によるアノテーションを行うことになる。濁点の付与もこの段階で行うことが望ましい。歴史的資料では，濁点が十分に付与されていないものが少なくないが，そのままでは読みにくく検索にとって不都合である。さらに，形態素解析を行う場合には辞書側での対応が必要になる上，曖昧性が増すため解析精度を低下させてしまう。

図 1 は，以上のアノテーションを施した状態の『明六雑誌コーパス』冒頭である。網掛け部分が濁点を付与した部分である。こうしたタグ付けは原則として人手で行わなければならない。

```
<magazine title="明六雑誌" year="1874" issue="01">
<front><titleBlock><block>
<s><pb n="1" originalN="1オ"/><lb/> 明六社雑誌第一號 </s>
</block></titleBlock></front>
<body>
<article title="洋字を以て国語を書するの論" author="西周" style="文語" script="漢字カタカナ">
<block><s><lb/>　洋字を以て國語を書するの論 </s></block>
<block><s>　西周 </s></block>
<p><s>
<lb/> 吾輩日常二三朋友の盍簪に於て偶當 <g type="包摂"> 時 </g> 治亂盛衰の故政治得失の跡な <lb/><vMark> ど </vMark> 凡て世故に就て談論爰に及 <vMark> ぶ </vMark><g type="包摂"> 時 </g> は動もすれ <vMark> ば </vMark> かの歐洲諸國と比較 <lb/> する 刁の多かる中に終には彼の文明を <g type="外字" ref="U+7FA1"> 羨 </g> み我 <vMark> が </vMark> 不開化を歎 <vMark> じ </vMark> 果て <odoriji originalText="々"> 果て </odoriji> は <lb/> 人民の愚如何ともするなしと云ふ 刁に歸して亦欷歔長大息に堪 <vMark> ざ </vMark> る <lb/> 者あり </s>
<s> 夫維新以來賢材も輩出し百度も更張し官省寮司より六十餘縣 <lb/> に至るま <vMark> で </vMark> 既に昔日の日本に非 <vMark> ず </vMark></s>
<s> 其 <g type="包摂"> 善 </g> 政美擧も屈指に暇あら <vMark> ざ </vMark> るな <lb/> り </s>
<s> 然るに退て熟々之を考ふれ <vMark> ば </vMark> 百端未 <vMark> だ </vMark> 脱垢の地に至ら <vMark> ざ </vMark> る事のみ <lb/> にして <g type="包摂"> 善 </g> 政あれ <vMark> ど </vMark> も民其澤を蒙ら <vMark> ず </vMark> 美擧あれ <vMark> ど </vMark> も得失相償は <vMark> ざ </vMark> る等の事 <lb/> 多し </s>
```

図 1　明六雑誌コーパスのアノテーション（単語情報を除く）

以上のアノテーションを施した後，形態素解析によって単語情報（短単位）のアノテーションを行う。この処理は，定評ある形態素解析器 MeCab[3]（Kudo et al. 2004）と，筆者らが開発

[2] http://www.tei-c.org
[3] https://code.google.com/p/mecab/

してきた形態素解析用の辞書である中古和文UniDic, 近代文語UniDic[4]を用いる。これにより，コーパス構築にとって実用的な 96% 以上の精度で解析を行うことができる。この後，形態論情報データベース（小木曽・中村 2011）上で解析誤りを人手で修正し，研究に利用可能なコーパスとなる。さらに長単位の情報を付与する場合には，短単位のデータを元に Comainu[5]（小澤ほか 2011）を用いてアノテーションを行う。このほかにも，より高度なアノテーションとして，後述する係り受け情報などが考えられる。

3.　濁点の自動付与

前節で見たとおり，歴史的な資料のコーパス化においては，形態素解析の前処理として濁点を付与する作業が必要になる場合がある。特に，国語研究所で開発を進めている近代語のコーパスでは，底本として校訂済みの本文ではなく原典を用いるため，この作業が必須である。『太陽』『明六雑誌』等では濁点が全く付されていないわけではなく，部分的に不完全な形で付与されている。従来はこの濁点付与作業を人間による目視で行ってきたが，膨大なテキストの全体を確認する必要があるため，熟練した作業者によっても見落としが少なくない。この作業に機械処理を導入することができれば，コーパス構築の作業負担を軽減することができる。

そのために，統計的機械学習の方法に基づく濁点の自動付与の研究を行った（岡ほか 2013）。この研究では，資料中に存在する濁点の付く可能性のある文字（清濁曖昧文字：「か，き，く，け，こ，さ，し，す，せ，そ，た，ち，つ，て，と，は，ひ，ふ，へ，ほ，ゝ，〳〵」）に対し，それぞれ独立に，濁点文字に置き換えるべきか否かを分類器を用いて判定する。提案手法では，分類の素性として対象文字の周辺文字列の情報だけを使用し，周囲の単語境界や品詞の情報は使用しない。そのため，学習用のコーパスは濁点が付与されたテキストだけでよい。そこで学習用コーパスには，整備済みで濁点無表記文字を含まない『太陽コーパス』のデータを利用した。このコーパス中から，学習対象の文字とその左右 3 文字を合わせて学習事例を作成している。分類器には 2 値分類器 LIBLINEAR[6]（Rong-En Fan et al. 2008）の L1 正則化ロジスティック回帰を用い，濁点を付ける事例を正例，濁点を付けない事例を負例としている。分類の素性には図 2 に示すように，分類対象文字と，その左右 3 文字の範囲内の文字 n-gram の組みを使用した。各 n-gram には出現位置（分類対象文字からの相対位置）を添え字として設けており，各素性は，「その位置にその n-gram が現れたか否か」を表す 2 値素性となっている。

この手法により，近代の雑誌『国民之友』を対象にした評価で適合率約 96%，再現率約 98% での濁点付与を達成した。この精度は，コーパスへのアノテーション補助に十分実用可能なものであり，今後「通時コーパス」プロジェクトでのコーパスの構築に応用することが期待される。

[4] http://www2.ninjal.ac.jp/lrc/index.php?UniDic
[5] https://maro.ninjal.ac.jp/Comainu/
[6] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

判定対象文字

↓

彼邦に譲らさるへき大雑誌を發行せんと

文字 n-gram

	-3	-2	-1	0	+1	+2	+3
1-gram	ら	さ	る	へ	き	大	雑
2-gram	らさ	さる	るへ	へき	き大	大雑	
3-gram	らさる	さるへ	るへき	へき大	き大雑		

図 2　濁点自動付与のための学習で使用する素性

こうして開発された自動濁点付与プログラムは，文系研究者にも利用可能な使いやすいアプリケーション「AYTC」として公開している（図 3）。AYTC は Silverlight[7] アプリケーションとして開発されており，特定の OS やブラウザに依存せず，幅広い PC 環境で利用することが可能になっている（岡ほか 2012）。

図 3　濁点自動付与アプリケーション AYTC

[7] http://www.microsoft.com/ja-jp/silverlight/

AYTC を用いることで，次のようなタグ付けを自動で行うことができる。

- 濁点付与を行った場合：本文は濁点文字に置き換え，濁点の付いていない元の文字をタグ内の属性「原文」に残し，次のようにタグ付けする
  <AYTC 原文 =”か” 確信度 =”0.9”> が </AYTC>
- 濁点を付与しなかった場合：AYTC タグを付けるだけで，本文への変更は行わない。
  <AYTC 確信度 =”0.4”> か </AYTC>

自動修正結果が 100％正しいわけではないため，人手による最終的なチェックは必要であるが，作業に当たってはタグ付けされる確信度を参考にしながら注意すべき箇所を絞り込むことができるため，完全に人手に頼る場合に比べ大幅な負担軽減を行うことが可能となった。

4.　より高度なアノテーションにむけて

『明六雑誌コーパス』は，単語情報（短単位）のアノテーションまで行ったものを公開している。現状の『日本語歴史コーパス　平安時代編』も，平安時代の仮名文学作品について，短単位の単語情報付与まで行ったものを公開しているが，今後，本プロジェクトでの研究成果を踏まえて，文節・長単位解析まで行ったデータを公開する予定である。これにより『現代日本語書き言葉均衡コーパス』と同等の形態論情報を付与された本格的な古典語コーパス

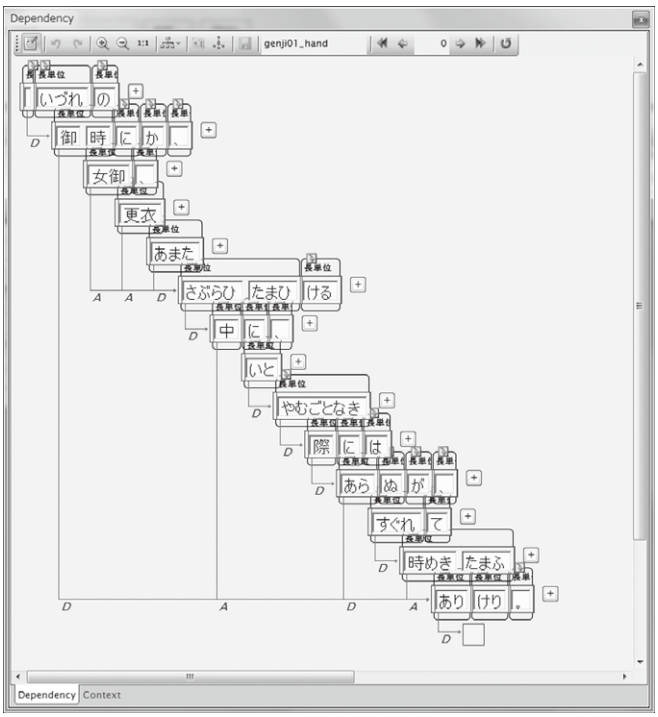

図 4　係り受け情報のアノテーション例（『源氏物語』冒頭）

が実現することになる。

しかし，限られた古典のデータを最大限に活かすためには，文節間の係り受けなどのより高度なアノテーションが期待される。たとえば，述語動詞にかかる要素がタグ付けされていれば，動詞の結合価（格パターン）などの情報を引き出すことができ，コーパスを利用した本格的なシンタクスの研究が可能になる。オックスフォード大学 VSARPJ プロジェクトによる Oxford Corpus of Old Japanese[8] では，すでに万葉集について限定的ながら句構造のマークアップを行っており，このような活用が可能になっている（ビャーケ・フレレスビッグ 2012）。当プロジェクトでは VSARPJ プロジェクトとも協力してこの問題について検討し，コーパス管理ツール「茶器」による環境（小木曽ほか 2011）を用意して係り受けアノテーションの試行を行ってきた（前ページの図 4）。

しかし，内省がきかない古典についてこのような高次のタグ付けを行うことにはたいへんな労力を必要とする。特に一文が長く係り先の曖昧性が高い散文では問題が大きい。『日本語歴史コーパス　平安時代編』は，万葉集と比較して分量もはるかに多く，大部分が散文であるため，係り受けのアノテーションには膨大な人手が必要となる。今後の歴史コーパス構築における課題の一つとして，必要性や応用可能性を考慮しつつ，実現の可能性を探っていきたいと考えている。

●参照文献●

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang and Chih-Jen Lin（2008）LIBLINEAR: A Library for Large Linear Classification, *Journal of Machine Learning Research* 9: 1871−1874.

フレレスビッグ，ビャーケ（2012）「オックスフォード上代日本語コーパスについて」『NINJAL「通時コーパス」プロジェクト・Oxford VSARPJ プロジェクト合同シンポジウム　通時コーパスと日本語史研究　予稿集』11−14.

国立国語研究所編（2005）『太陽コーパス』（国立国語研究所資料集 15）．東京：博文館新社.

近藤明日子・田中牧郎（2012）「『明六雑誌コーパス』の仕様」『近代語コーパス設計のための文献言語研究　成果報告書』（国立国語研究所共同研究報告 12-03），118−143.
http://www.ninjal.ac.jp/corpus_center/cmj/doc/07kondo.pdf

Kudo, Taku, Kaoru Yamamoto, and Yuji Matsumoto（2004）Applying conditional random fields to Japanese morphological analysis. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*（Barcelona, Spain），230−237.

小木曽智信・中村壮範（2011）『『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装　改訂版』（国立国語研究所内部報告書 LR-CCG-10-06）.

小木曽智信・岡照晃・小町守・松本裕治（2011）「コーパス管理ツール「茶器」による単語情報付き古典語コーパスの活用」『人文科学とコンピュータシンポジウム「じんもんこん 2011」』2011（8）: 255−260.

岡照晃・小町守・小木曽智信・松本裕治（2012）「未整備の歴史的文献への濁点の自動付与アプリケーション」『人文科学とコンピュータシンポジウム「じんもんこん 2012」』2012（7）: 191−198.

岡照晃・小町守・小木曽智信・松本裕治（2013）「統計的機械学習を用いた歴史的資料への濁点付与の自動化」『情報処理学会論文誌』54（4）: 1641−1654.

小澤俊介・内元清貴・伝康晴（2011）「BCCWJ に基づく中・長単位解析ツール」『特定領域「日本語コーパス」平成 22 年度公開ワークショップ予稿集』331−338.

[8] http://vsarpj.orinst.ox.ac.uk/corpus/

《要旨》 通時コーパスの構築に必要とされる歴史的日本語資料のアノテーションの全体について俯瞰した上で，アノテーション作業の自動化の試みの一つとして濁点の自動付与に関する研究成果を紹介する。歴史的資料では，濁点が十分に付与されていないものが少なくないが，そのままでは読みにくく検索や形態素解析にとって不都合である。そこで統計的機械学習に基づく自動濁点付与の手法を開発し，適合率約 96%，再現率約 98% での濁点付与を可能にした。これにより通時コーパス構築の作業負担の軽減が期待できる。最後に，今後の歴史コーパスに期待される高度なアノテーションについて展望する。

Abstract: Following a survey of annotations for historical Japanese documents that are required for the construction of a diachronic corpus, I introduce the results of our research on adding *dakuten* (the voicing diacritic) automatically. Raw historical texts often include characters with *dakuten* omitted, but such texts degrade readability and retrievability and are not suitable for morphological analysis. We therefore developed an automatic annotation technique for *dakuten* based on statistical machine learning that has a precision rate of approximately 96% and a recall rate of approximately 98%. This technique can reduce the work involved in diachronic corpus construction. Finally, I discuss the high-level annotation that can be expected in diachronic corpora from now on.

小木曽 智信 (おぎそ・としのぶ)

国立国語研究所言語資源研究系准教授。修士（文学）。東京大学大学院人文社会系研究科博士課程単位取得満期退学。明海大学専任講師，独立行政法人国立国語研究所研究員を経て 2009 年 10 月より現職。コーパス開発センター兼任。「現代日本語書き言葉均衡コーパス」「日本語歴史コーパス」の構築に携わる。

主な著書・論文：『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―』（共著，博文館新社，2005），『講座日本語コーパス 1　コーパス入門』（共著，朝倉書店，2013 年）．

萌芽・発掘型共同研究プロジェクト「統計と機械学習による日本語史研究」
プロジェクトリーダー　小木曽智信
（国立国語研究所 言語資源研究系 准教授）

プロジェクトの概要

自然言語処理の技術が発展し，電子化辞書の整備が進んだことにより，従来は不可能であった歴史的資料を対象とした形態素解析が可能になった。これにより日本語史の分野においてもコーパスと統計的手法を活用した新しいタイプの研究が可能になりつつある。

本プロジェクトでは，機械学習の手法を用いて日本語通時コーパスの整備に必要となる各種の技術を開発し，多様な日本語史資料に対する高度なアノテーションを可能にする。同時に，既存のツールを応用して日本語史研究のためのコーパス利用環境を整備する。そして整備したコーパスとその利用環境を用いて，多変量解析などの統計的手法に基づく新しい方法による日本語史研究に取り組む。

開発したソフトウェアと研究成果は一般に公開するとともに，国語研で計画中の通時コーパスの構築に活用する。